RNA-Seq Mini Project

Hayoung A15531571

11/18/2021

Differential Expression Analysis

First we load our package

```
library(DESeq2)
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
## Loading required package: GenomicRanges
## Loading required package: GenomeInfoDb
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
##
## Attaching package: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

Then let's load our files in

```r
metaFile <- "GSE37704_metadata.csv"
countFile <- "GSE37704_featurecounts.csv"
```

```r
# Import metadata and take a peak
colData <- read.csv(metaFile, row.names=1)
head(colData)
```

```
##               condition
## SRR493366 control_sirna
## SRR493367 control_sirna
## SRR493368 control_sirna
## SRR493369      hoxa1_kd
## SRR493370      hoxa1_kd
## SRR493371      hoxa1_kd
```

```r
# Import countdata
countData <- read.csv(countFile, row.names=1)
head(countData)
```

```
##                 length SRR493366 SRR493367 SRR493368 SRR493369 SRR493370
## ENSG00000186092    918         0         0         0         0         0
## ENSG00000279928    718         0         0         0         0         0
## ENSG00000279457   1982        23        28        29        29        28
## ENSG00000278566    939         0         0         0         0         0
## ENSG00000273547    939         0         0         0         0         0
## ENSG00000187634   3214       124       123       205       207       212
##                 SRR493371
## ENSG00000186092         0
## ENSG00000279928         0
## ENSG00000279457        46
## ENSG00000278566         0
## ENSG00000273547         0
## ENSG00000187634       258
```

Q. Complete the code below to remove the troublesome first column from countData

```r
# Note we need to remove the odd first $length col
countData <- as.matrix(countData[,-1])
head(countData)
```

```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000186092         0         0         0         0         0         0
## ENSG00000279928         0         0         0         0         0         0
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000278566         0         0         0         0         0         0
## ENSG00000273547         0         0         0         0         0         0
## ENSG00000187634       124       123       205       207       212       258
```

Q. Complete the code below to filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
# Filter count data where you have 0 read count across all samples.
zero.vals <- which(countData[,1:2]==0, arr.ind=TRUE)

to.rm <- unique(zero.vals[,1])
countData <- countData[-to.rm,]
head(countData)
```

```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000187634       124       123       205       207       212       258
## ENSG00000188976      1637      1831      2383      1226      1326      1504
## ENSG00000187961       120       153       180       236       255       357
## ENSG00000187583        24        48        65        44        48        64
## ENSG00000187642         4         9        16        14        16        16
```

#Running DESeq2

```
#Setup the object
dds = DESeqDataSetFromMatrix(countData=countData,
                             colData=colData,
                             design=~condition)
```

```
## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

```
#Run it
dds = DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
#Get our results
res <- results(dds)
head(res)
```

```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 6 rows and 6 columns
##                  baseMean log2FoldChange     lfcSE       stat      pvalue
##                 <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1802410 0.3128743   0.576081 5.64560e-01
## ENSG00000187634  183.2296      0.4259300 0.1357991   3.136471 1.70994e-03
## ENSG00000188976 1651.1881     -0.6927121 0.0549826 -12.598761 2.14486e-36
## ENSG00000187961  209.6379      0.7299474 0.1279936   5.702998 1.17718e-08
## ENSG00000187583   47.2551      0.0393402 0.2613090   0.150550 8.80330e-01
## ENSG00000187642   11.9798      0.5397049 0.5013479   1.076508 2.81700e-01
##                        padj
##                   <numeric>
## ENSG00000279457 6.53784e-01
## ENSG00000187634 3.52201e-03
## ENSG00000188976 2.40943e-35
## ENSG00000187961 4.06810e-08
## ENSG00000187583 9.12748e-01
## ENSG00000187642 3.68486e-01
```

Next, get results for the HoxA1 knockdown versus control siRNA

```
res = results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

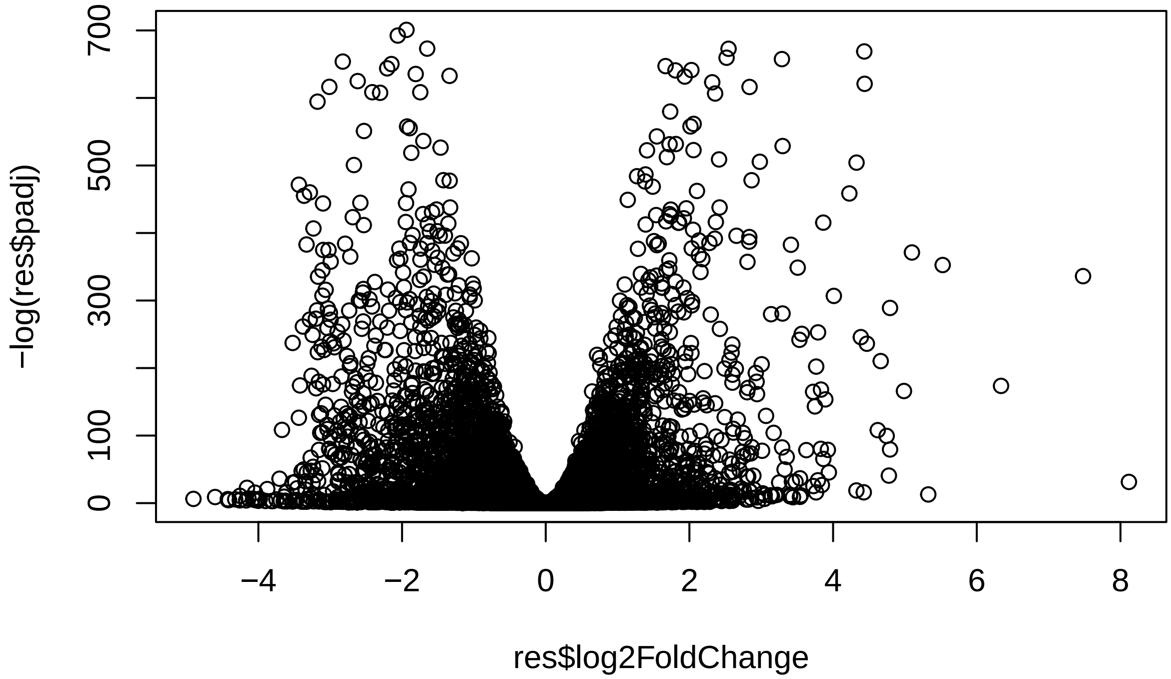
> Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

```
summary(res)
```

```
##
## out of 13761 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 4328, 31%
## LFC < 0 (down)     : 4474, 33%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

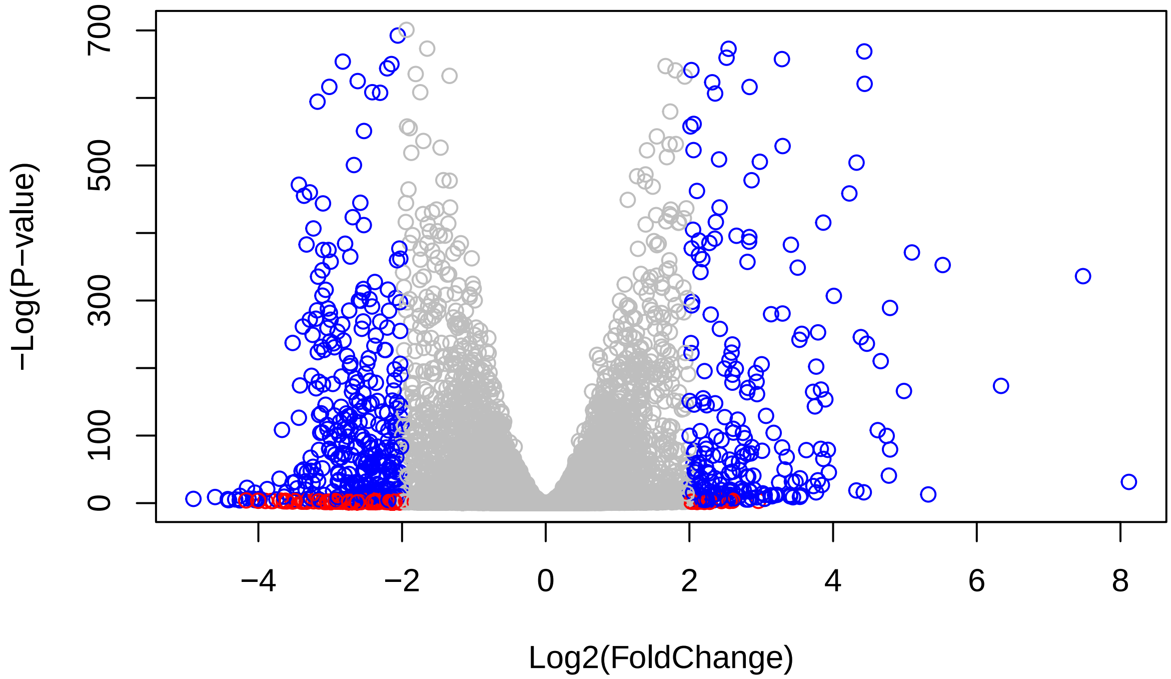
#Volcano Plot

```
plot( res$log2FoldChange, -log(res$padj) )
```



Q. Improve this plot by completing the below code, which adds color and axis labels

```
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )

# Color red the genes with absolute fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"

# Color blue those with adjusted p-value less than 0.01
#  and absolute fold change more than 2
inds <- (abs(res$pvalue) < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P-value)" )
```



Adding Gene Annotation

Q. Use the mapIDs() function multiple times to add SYMBOL, ENTREZID and GENENAME annotation to our results by completing the code below.

```
library("AnnotationDbi")
```

```
## Warning: package 'AnnotationDbi' was built under R version 4.1.2
```

```
library("org.Hs.eg.db")
```

```
##
```

```
columns(org.Hs.eg.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
##  [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "MAP"
## [16] "OMIM"         "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
## [21] "PMID"         "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
## [26] "UNIPROT"
```

```
res$symbol = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="SYMBOL",
                    multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
res$entrez = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="ENTREZID",
                    multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
res$name =   mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="GENENAME",
                    multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
head(res, 10)
```

```
## log2 fold change (MLE): condition hoxa1_kd vs control_sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 10 rows and 9 columns
##                  baseMean log2FoldChange     lfcSE      stat      pvalue
##                 <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1802410 0.3128743   0.576081 5.64560e-01
## ENSG00000187634  183.2296      0.4259300 0.1357991   3.136471 1.70994e-03
## ENSG00000188976 1651.1881     -0.6927121 0.0549826 -12.598761 2.14486e-36
## ENSG00000187961  209.6379      0.7299474 0.1279936   5.702998 1.17718e-08
## ENSG00000187583   47.2551      0.0393402 0.2613090   0.150550 8.80330e-01
## ENSG00000187642   11.9798      0.5397049 0.5013479   1.076508 2.81700e-01
## ENSG00000188290  108.9221      2.0563306 0.1914001  10.743624 6.35019e-27
## ENSG00000187608  350.7169      0.2570463 0.1001328   2.567054 1.02567e-02
## ENSG00000188157 9128.4394      0.3899096 0.0481440   8.098821 5.54943e-16
## ENSG00000131591  156.4791      0.1968739 0.1409590   1.396675 1.62511e-01
##                        padj      symbol      entrez                   name
##                   <numeric> <character> <character>            <character>
## ENSG00000279457 6.53784e-01      WASH9P   102723897 WAS protein family h..
## ENSG00000187634 3.52201e-03      SAMD11      148398 sterile alpha motif ..
## ENSG00000188976 2.40943e-35       NOC2L       26155 NOC2 like nucleolar ..
## ENSG00000187961 4.06810e-08      KLHL17      339451 kelch like family me..
## ENSG00000187583 9.12748e-01     PLEKHN1       84069 pleckstrin homology ..
## ENSG00000187642 3.68486e-01       PERM1       84808 PPARGC1 and ESRR ind..
## ENSG00000188290 5.26099e-26        HES4       57801 hes family bHLH tran..
## ENSG00000187608 1.87489e-02       ISG15        9636 ISG15 ubiquitin like..
## ENSG00000188157 2.94735e-15        AGRN      375790                  agrin
## ENSG00000131591 2.29875e-01    C1orf159       54991 chromosome 1 open re..
```

```
columns(org.Hs.eg.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
##  [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "MAP"
## [16] "OMIM"         "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
## [21] "PMID"         "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
## [26] "UNIPROT"
```

> Q. Finally for this section let's reorder these results by adjusted p-value and save them to a CSV file in your current project directory.

```
res = res[order(res$pvalue),]
write.csv(res, file="deseq_results.csv")
```

Pathway Analysis

Load them in

```
library(pathview)
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################
```

```
library(gage)
```

```
##
```

```
library(gageData)
```

```
data(kegg.sets.hs)
data(sigmet.idx.hs)

# Focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

# Examine the first 3 pathways
head(kegg.sets.hs, 3)
```

```
## $‘hsa00232 Caffeine metabolism‘
## [1] "10"   "1544" "1548" "1549" "1553" "7498" "9"   
## 
## $‘hsa00983 Drug metabolism - other enzymes‘
##  [1] "10"     "1066"   "10720"  "10941"  "151531" "1548"   "1549"   "1551"  
##  [9] "1553"   "1576"   "1577"   "1806"   "1807"   "1890"   "221223" "2990"  
## [17] "3251"   "3614"   "3615"   "3704"   "51733"  "54490"  "54575"  "54576" 
## [25] "54577"  "54578"  "54579"  "54600"  "54657"  "54658"  "54659"  "54963" 
## [33] "574537" "64816"  "7083"   "7084"   "7172"   "7363"   "7364"   "7365"  
## [41] "7366"   "7367"   "7371"   "7372"   "7378"   "7498"   "79799"  "83549" 
## [49] "8824"   "8833"   "9"      "978"   
## 
## $‘hsa00230 Purine metabolism‘
##   [1] "100"    "10201"  "10606"  "10621"  "10622"  "10623"  "107"    "10714" 
##   [9] "108"    "10846"  "109"    "111"    "11128"  "11164"  "112"    "113"   
##  [17] "114"    "115"    "122481" "122622" "124583" "132"    "158"    "159"   
##  [25] "1633"   "171568" "1716"   "196883" "203"    "204"    "205"    "221823"
##  [33] "2272"   "22978"  "23649"  "246721" "25885"  "2618"   "26289"  "270"   
##  [41] "271"    "27115"  "272"    "2766"   "2977"   "2982"   "2983"   "2984"  
##  [49] "2986"   "2987"   "29922"  "3000"   "30833"  "30834"  "318"    "3251"  
##  [57] "353"    "3614"   "3615"   "3704"   "377841" "471"    "4830"   "4831"  
##  [65] "4832"   "4833"   "4860"   "4881"   "4882"   "4907"   "50484"  "50940" 
##  [73] "51082"  "51251"  "51292"  "5136"   "5137"   "5138"   "5139"   "5140"  
##  [81] "5141"   "5142"   "5143"   "5144"   "5145"   "5146"   "5147"   "5148"  
##  [89] "5149"   "5150"   "5151"   "5152"   "5153"   "5158"   "5167"   "5169"  
##  [97] "51728"  "5198"   "5236"   "5313"   "5315"   "53343"  "54107"  "5422"  
## [105] "5424"   "5425"   "5426"   "5427"   "5430"   "5431"   "5432"   "5433"  
## [113] "5434"   "5435"   "5436"   "5437"   "5438"   "5439"   "5440"   "5441"  
## [121] "5471"   "548644" "55276"  "5557"   "5558"   "55703"  "55811"  "55821" 
## [129] "5631"   "5634"   "56655"  "56953"  "56985"  "57804"  "58497"  "6240"  
## [137] "6241"   "64425"  "646625" "654364" "661"    "7498"   "8382"   "84172" 
## [145] "84265"  "84284"  "84618"  "8622"   "8654"   "87178"  "8833"   "9060"  
## [153] "9061"   "93034"  "953"    "9533"   "954"    "955"    "956"    "957"   
## [161] "9583"   "9615"
```

```
foldchanges = res$log2FoldChange
names(foldchanges) = res$entrez
head(foldchanges)
```

```
##      1266     54855      1465      2034      2150      6659 
## -2.422685  3.201862 -2.313714 -1.888000  3.344481  2.392259
```

Gage pathway analysis

```
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)
```

```
attributes(keggres)
```

```
## $names
## [1] "greater" "less"    "stats"
```

```
# Look at the first few down (less) pathways
head(keggres$less)
```

```
##                                          p.geomean stat.mean        p.val
## hsa04110 Cell cycle                   1.888472e-05 -4.205434 1.888472e-05
## hsa03030 DNA replication              1.209058e-04 -3.871120 1.209058e-04
## hsa04114 Oocyte meiosis               7.921929e-04 -3.206473 7.921929e-04
## hsa03440 Homologous recombination     4.227051e-03 -2.734017 4.227051e-03
## hsa00010 Glycolysis / Gluconeogenesis 6.053365e-03 -2.563476 6.053365e-03
## hsa00240 Pyrimidine metabolism        1.172151e-02 -2.285838 1.172151e-02
##                                            q.val set.size         exp1
## hsa04110 Cell cycle                   0.002964901      119 1.888472e-05
## hsa03030 DNA replication              0.009491108       36 1.209058e-04
## hsa04114 Oocyte meiosis               0.041458097       95 7.921929e-04
## hsa03440 Homologous recombination     0.165911753       28 4.227051e-03
## hsa00010 Glycolysis / Gluconeogenesis 0.190075653       44 6.053365e-03
## hsa00240 Pyrimidine metabolism        0.283903993       90 1.172151e-02
```

Download some pictures of the pathways

```
pathview(gene.data=foldchanges, pathway.id="hsa04110")
```

```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04110.pathview.png
```

```
# A different PDF based output of the same data
pathview(gene.data=foldchanges, pathway.id="hsa04110", kegg.native=FALSE)
```

```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04110.pathview.pdf
```

```
## Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$greater)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
keggresids
```

```
## [1] "hsa04142" "hsa04640" "hsa04630" "hsa04380" "hsa00140"
```

```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

```
## Info: Downloading xml files for hsa04142, 1/1 pathways..
## Info: Downloading png files for hsa04142, 1/1 pathways..

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04142.pathview.png

## Info: some node width is different from others, and hence adjusted!

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04640.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04630.pathview.png

## Info: Downloading xml files for hsa04380, 1/1 pathways..

## Info: Downloading png files for hsa04380, 1/1 pathways..

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04380.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa00140.pathview.png
```

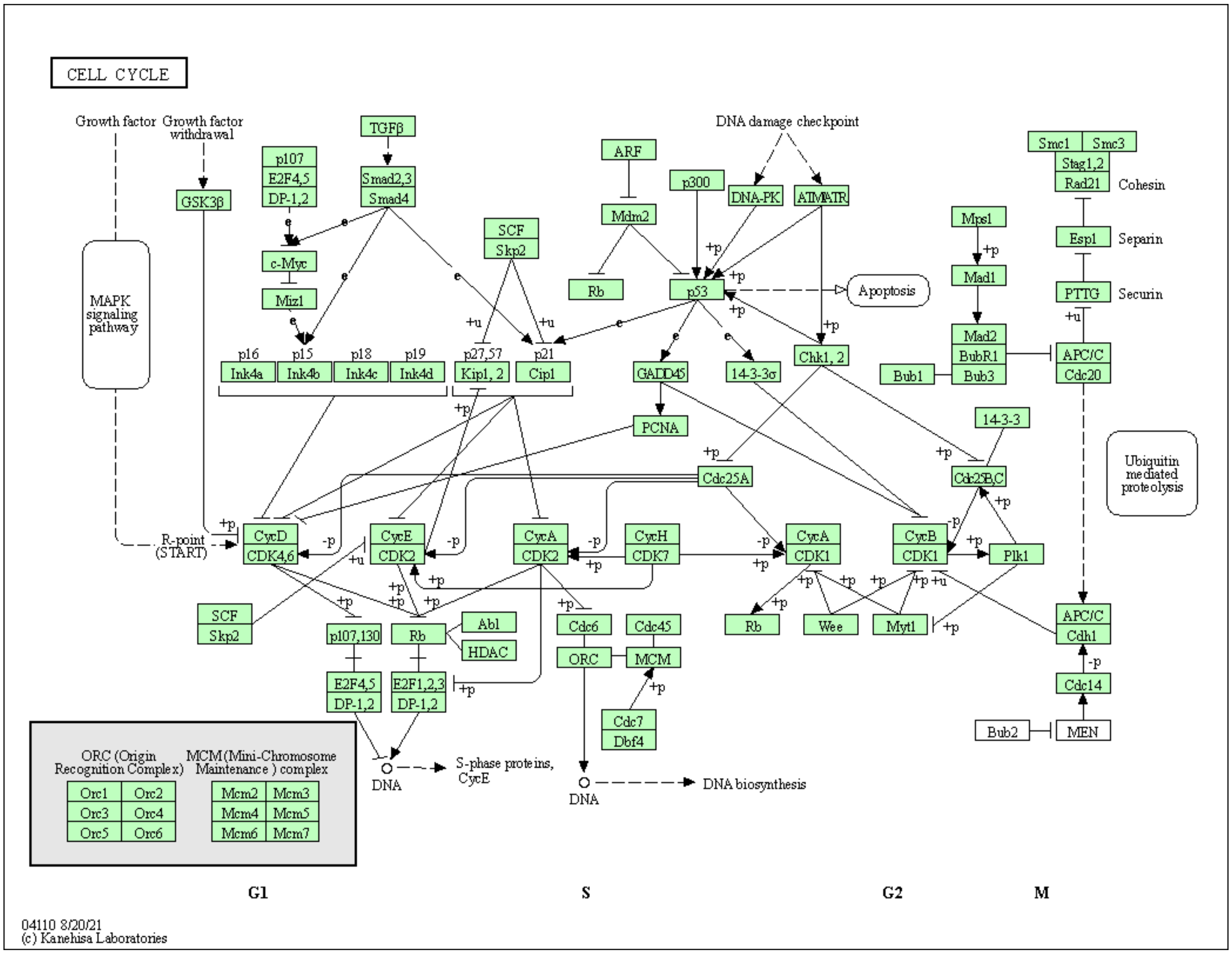
> Q. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

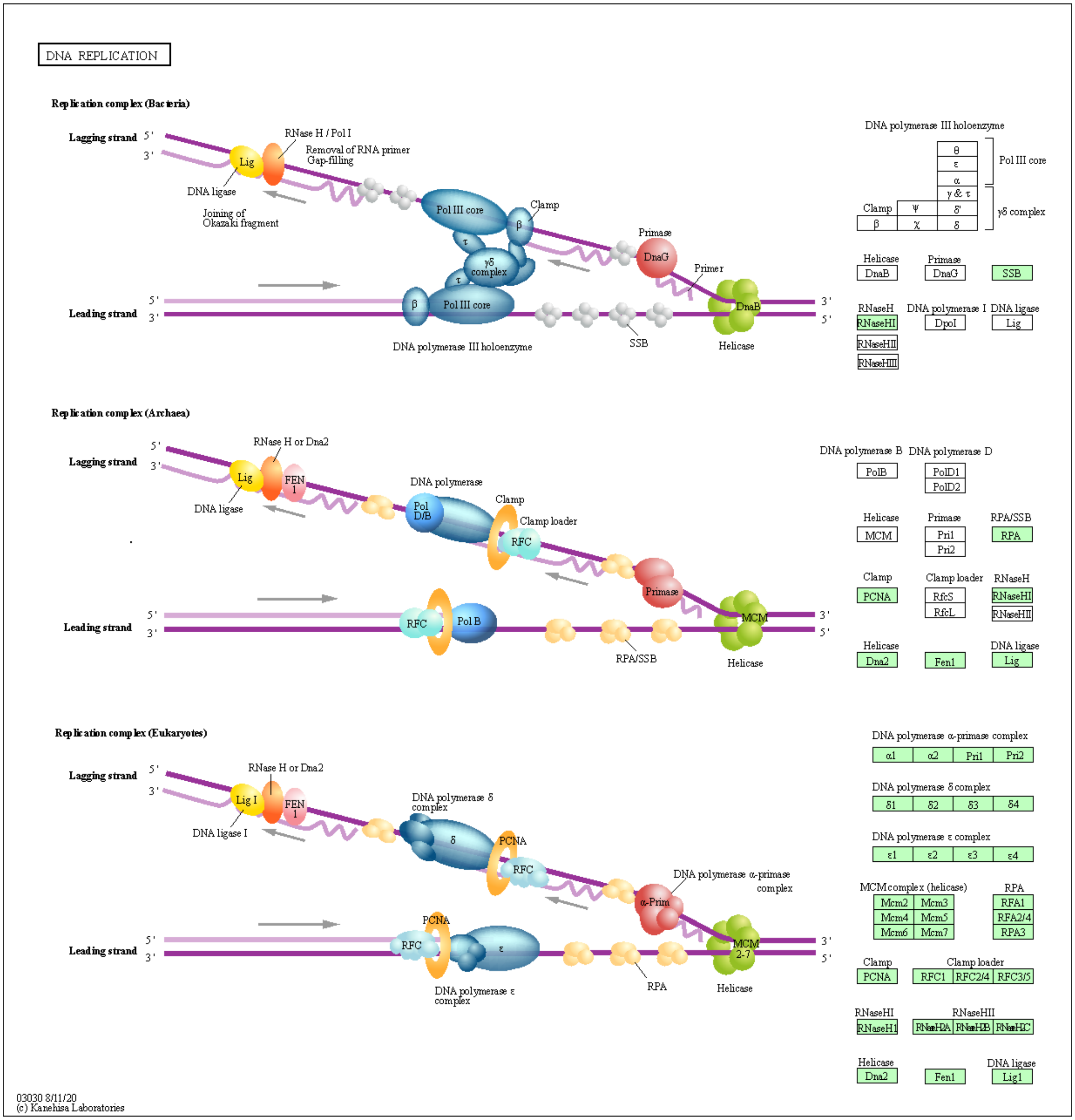
```r
## Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$less)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
keggresids
```

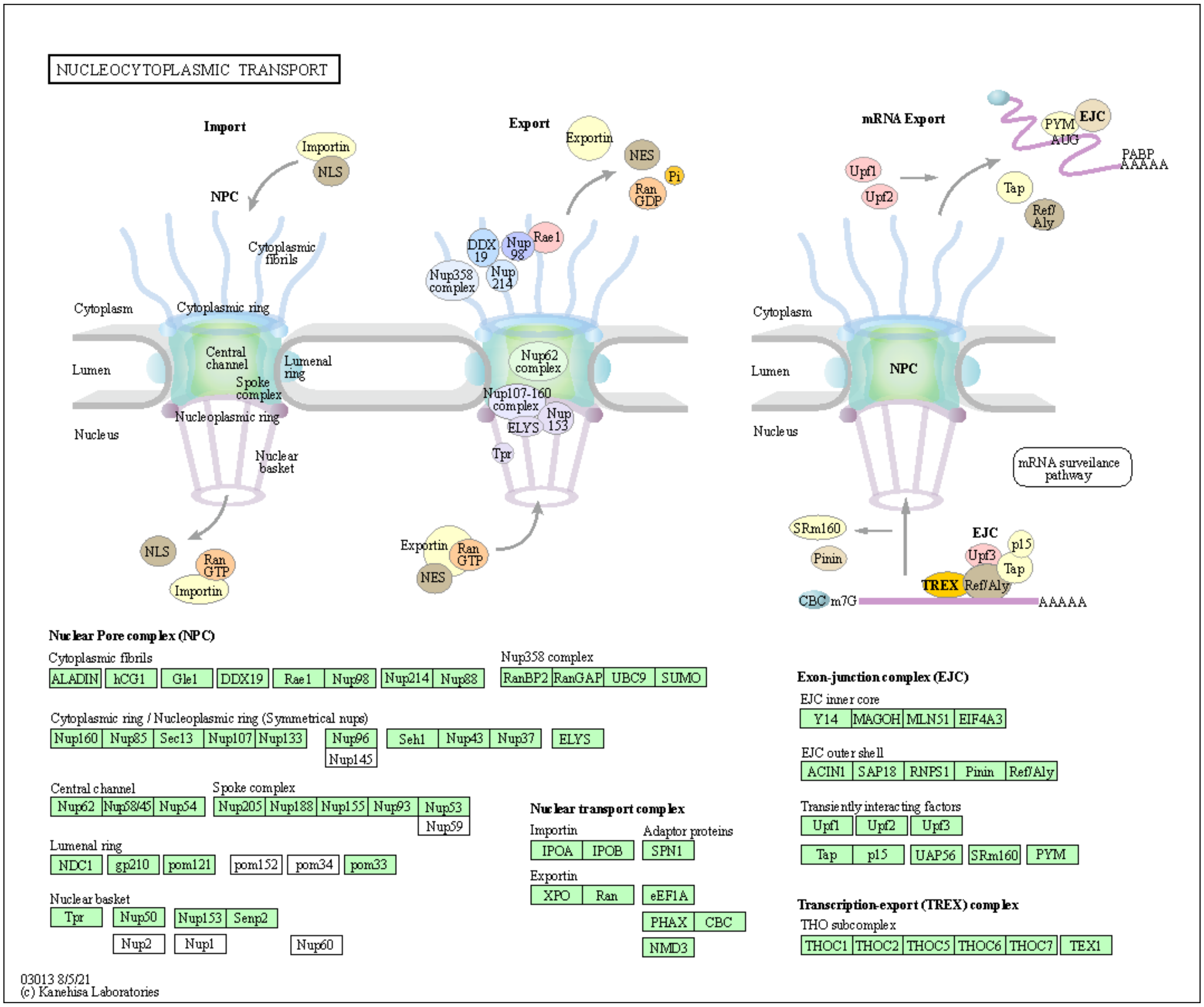
```
## [1] "hsa04110" "hsa03030" "hsa04114" "hsa03440" "hsa00010"
```

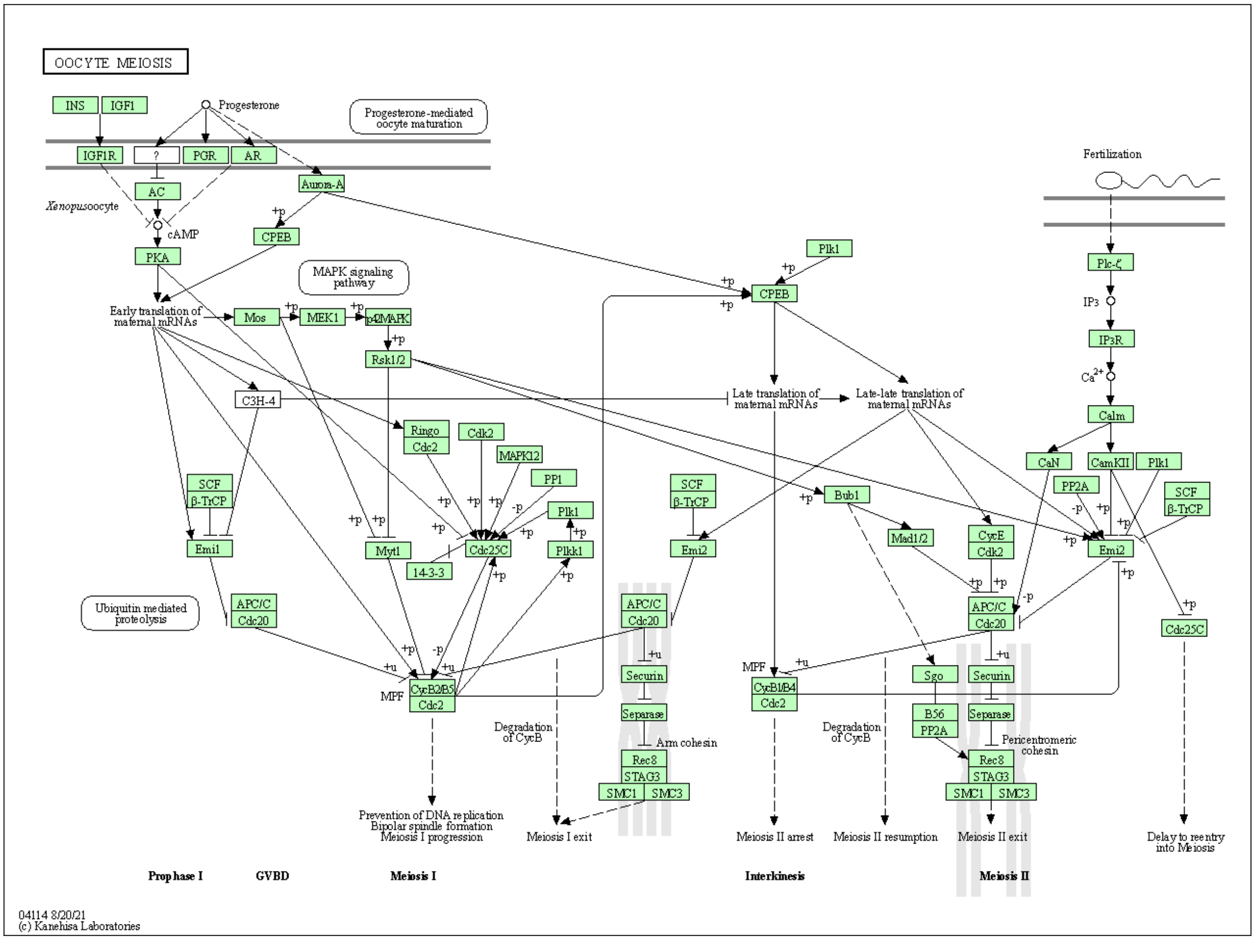
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

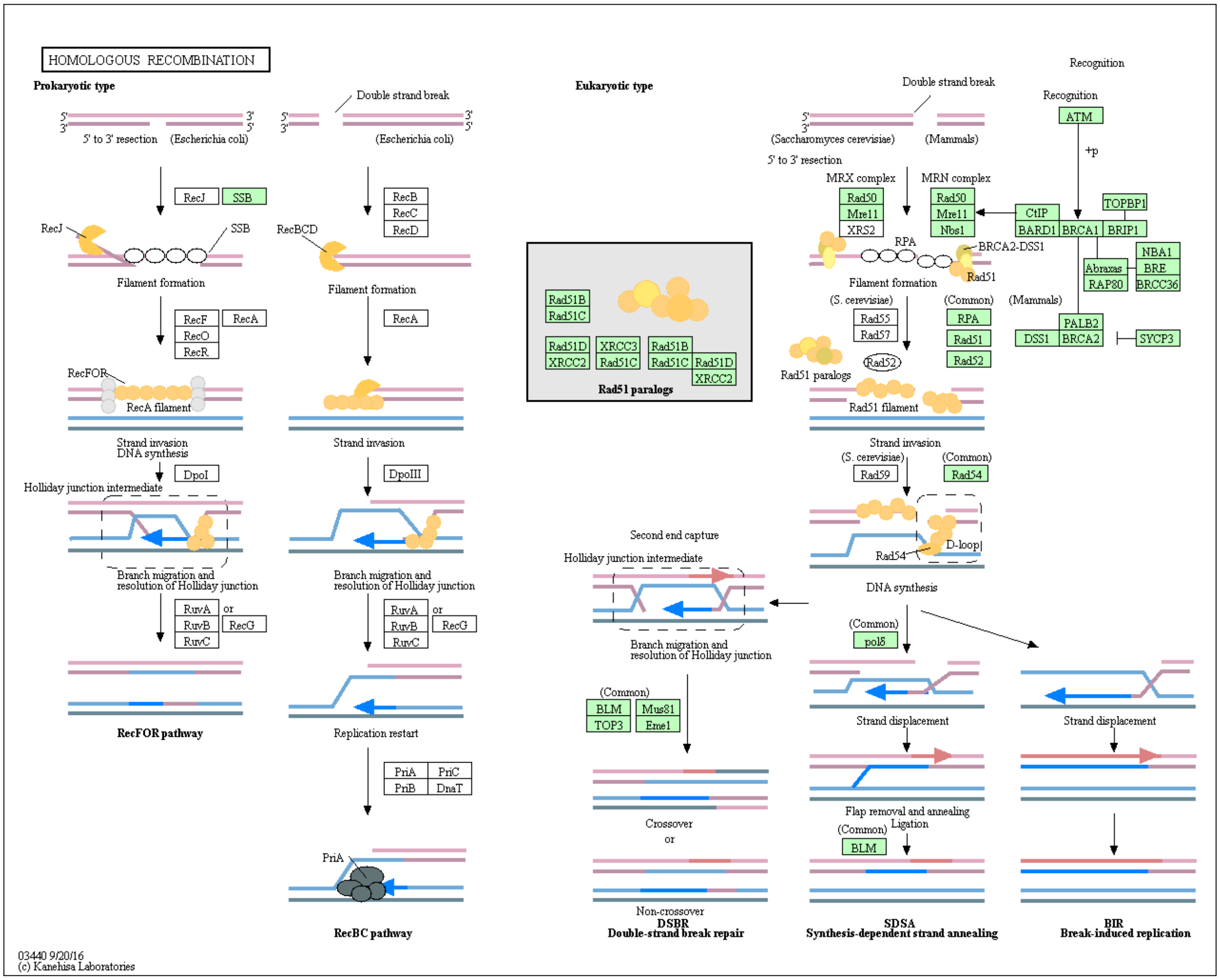
```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04110.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa03030.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa04114.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa03440.pathview.png

## Info: Downloading xml files for hsa00010, 1/1 pathways..

## Info: Downloading png files for hsa00010, 1/1 pathways..

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/hayoungpark/Desktop/bimm143 class/github stuff/githubs/Class16_Min

## Info: Writing image file hsa00010.pathview.png
```


CELL CYCLE
Growth factor
Growth factor withdrawal
GSK3β
MAPK signaling pathway
R-point (START)
p107
E2F4,5
DP-1,2
TGFβ
Smad2,3
Smad4
c-Myc
Miz1
SCF
Skp2
ARF
Mdm2
Rb
p300
p53
DNA damage checkpoint
DNA-PK
ATM/ATR
Apoptosis
Chk1, 2
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Esp1
Separin
PTTG
Securin
Mps1
Mad1
Mad2
BubR1
Bub3
Bub1
APC/C
Cdc20
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
GADD45
14-3-3σ
PCNA
14-3-3
Cdc25B,C
Cdc25A
Ubiquitin mediated proteolysis
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
Cdc6
Cdc45
ORC
MCM
Rb
Wee
Myt1
APC/C
Cdh1
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc7
Dbf4
Cdc14
Bub2
MEN
ORC (Origin Recognition Complex)
MCM (Mini-Chromosome Maintenance) complex
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
+p
-p
+u
e
G1
S
G2
M
04110 8/20/21
(c) Kanehisa Laboratories


DNA REPLICATION
Replication complex (Bacteria)
Lagging strand
5'
3'
Lig
RNase H / Pol I
Removal of RNA primer
Gap-filling
DNA ligase
Joining of
Okazaki fragment
Pol III core
Clamp
β
τ
γδ
complex
τ
Primase
DnaG
Primer
DnaB
Leading strand
5'
3'
β
Pol III core
3'
5'
DNA polymerase III holoenzyme
SSB
Helicase
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
Pol III core
Clamp
ψ
δ'
β
χ
δ
γδ complex
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
Lagging strand
5'
3'
RNase H or Dna2
Lig
FEN
1
DNA ligase
DNA polymerase
Pol
D/B
Clamp
Clamp loader
RFC
Primase
MCM
Leading strand
5'
3'
RFC
Pol B
3'
5'
RPA/SSB
Helicase
DNA polymerase B
PolB
DNA polymerase D
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
RNaseHI
RNaseHII
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
Lagging strand
5'
3'
RNase H or Dna2
Lig I
FEN
1
DNA ligase I
DNA polymerase δ
complex
δ
PCNA
RFC
DNA polymerase α-primase
complex
α-Prim
MCM
2-7
PCNA
Leading strand
5'
3'
RFC
ε
3'
5'
DNA polymerase ε
complex
RPA
Helicase
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RFA2/4
RPA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseH1
RNaseHII
RNaseH2A
RNaseH2B
RNaseH2C
Helicase
Dna2
Fen1
DNA ligase
Lig1
03030 8/11/20
(c) Kanehisa Laboratories


NUCLEOCYTOPLASMIC TRANSPORT
Import
Importin
NLS
NPC
Cytoplasmic fibrils
Cytoplasm
Cytoplasmic ring
Lumen
Central channel
Lumenal ring
Spoke complex
Nucleoplasmic ring
Nucleus
Nuclear basket
NLS
Ran GTP
Importin
Export
Exportin
NES
Pi
Ran GDP
Rae1
DDX 19
Nup 98
Nup358 complex
Nup 214
Nup62 complex
Nup107-160 complex
Nup 153
ELYS
Tpr
Exportin
Ran GTP
NES
mRNA Export
Upf1
Upf2
PYM
EJC
AUG
PABP AAAAA
Tap
Ref/ Aly
Cytoplasm
Lumen
NPC
Nucleus
mRNA surveilance pathway
SRm160
Pinin
EJC
Upf3
p15
Tap
TREX
Ref/Aly
CBC
m7G
AAAAA
Nuclear Pore complex (NPC)
Cytoplasmic fibrils
ALADIN
hCG1
Gle1
DDX19
Rae1
Nup98
Nup214
Nup88
Nup358 complex
RanBP2
RanGAP
UBC9
SUMO
Cytoplasmic ring / Nucleoplasmic ring (Symmetrical nups)
Nup160
Nup85
Sec13
Nup107
Nup133
Nup96
Nup145
Seh1
Nup43
Nup37
ELYS
Central channel
Nup62
Nup58/45
Nup54
Spoke complex
Nup205
Nup188
Nup155
Nup93
Nup53
Nup59
Lumenal ring
NDC1
gp210
pom121
pom152
pom34
pom33
Nuclear basket
Tpr
Nup50
Nup153
Senp2
Nup2
Nup1
Nup60
Nuclear transport complex
Importin
IPOA
IPOB
Adaptor proteins
SPN1
Exportin
XPO
Ran
eEF1A
PHAX
CBC
NMD3
Exon-junction complex (EJC)
EJC inner core
Y14
MAGOH
MLN51
EIF4A3
EJC outer shell
ACIN1
SAP18
RNPS1
Pinin
Ref/Aly
Transiently interacting factors
Upf1
Upf2
Upf3
Tap
p15
UAP56
SRm160
PYM
Transcription-export (TREX) complex
THO subcomplex
THOC1
THOC2
THOC5
THOC6
THOC7
TEX1
03013 8/5/21
(c) Kanehisa Laboratories


OOCYTE MEIOSIS
INS
IGF1
Progesterone
Progesterone-mediated oocyte maturation
IGF1R
?
PGR
AR
AC
Aurora-A
Xenopus oocyte
cAMP
+p
CPEB
PKA
MAPK signaling pathway
Early translation of maternal mRNAs
Mos
+p
MEK1
+p
p42MAPK
+p
Rsk1/2
C3H-4
Ringo
Cdc2
Cdk2
MAPK12
PP1
Plk1
SCF
β-TrCP
Emi1
Myt1
14-3-3
Cdc25C
Plkk1
Ubiquitin mediated proteolysis
APC/C
Cdc20
MPF
CycB2/B5
Cdc2
Degradation of CycB
Prevention of DNA replication
Bipolar spindle formation
Meiosis I progression
Meiosis I exit
APC/C
Cdc20
Securin
Separase
Arm cohesin
Rec8
STAG3
SMC1
SMC3
Plk1
CPEB
SCF
β-TrCP
Emi2
Late translation of maternal mRNAs
Late-late translation of maternal mRNAs
Bub1
Mad1/2
CycE
Cdk2
APC/C
Cdc20
MPF
CycB1/B4
Cdc2
Degradation of CycB
Sgo
Securin
B56
PP2A
Separase
Pericentromeric cohesin
Rec8
STAG3
SMC1
SMC3
Meiosis II arrest
Meiosis II resumption
Meiosis II exit
Fertilization
Plc-ζ
IP3
IP3R
Ca2+
Calm
CaN
CamKII
Plk1
PP2A
SCF
β-TrCP
Emi2
Cdc25C
Delay to reentry into Meiosis
+p
-p
+u
Prophase I
GVBD
Meiosis I
Interkinesis
Meiosis II
04114 8/20/21
(c) Kanehisa Laboratories



Gene Ontology

```
data(go.sets.hs)
data(go.subs.hs)

# Focus on Biological Process subset of GO
gobpsets = go.sets.hs[go.subs.hs$BP]

gobpres = gage(foldchanges, gsets=gobpsets, same.dir=TRUE)

lapply(gobpres, head)
```

```
## $greater
##                                                          p.geomean stat.mean
## GO:0007156 homophilic cell adhesion                   3.574409e-05  4.065745
## GO:0016339 calcium-dependent cell-cell adhesion       6.624322e-04  3.414326
## GO:0048729 tissue morphogenesis                       9.629642e-04  3.113452
## GO:0002009 morphogenesis of an epithelium             1.036665e-03  3.093930
## GO:1901617 organic hydroxy compound biosynthetic process 1.825666e-03  2.937016
## GO:0035295 tube development                              2.137116e-03  2.867380
##                                                                 p.val     q.val
## GO:0007156 homophilic cell adhesion                      3.574409e-05 0.1348982
## GO:0016339 calcium-dependent cell-cell adhesion          6.624322e-04 0.6085845
## GO:0048729 tissue morphogenesis                          9.629642e-04 0.6085845
## GO:0002009 morphogenesis of an epithelium                1.036665e-03 0.6085845
## GO:1901617 organic hydroxy compound biosynthetic process 1.825666e-03 0.6085845
## GO:0035295 tube development                              2.137116e-03 0.6085845
##                                                          set.size         exp1
## GO:0007156 homophilic cell adhesion                            91 3.574409e-05
## GO:0016339 calcium-dependent cell-cell adhesion                25 6.624322e-04
## GO:0048729 tissue morphogenesis                               356 9.629642e-04
## GO:0002009 morphogenesis of an epithelium                     289 1.036665e-03
## GO:1901617 organic hydroxy compound biosynthetic process      119 1.825666e-03
## GO:0035295 tube development                                   335 2.137116e-03
##
## $less
##                                             p.geomean stat.mean        p.val
## GO:0000279 M phase                       1.070282e-15 -8.081854 1.070282e-15
## GO:0048285 organelle fission             1.486831e-14 -7.771854 1.486831e-14
## GO:0000280 nuclear division              2.849163e-14 -7.694716 2.849163e-14
## GO:0007067 mitosis                       2.849163e-14 -7.694716 2.849163e-14
## GO:0000087 M phase of mitotic cell cycle 9.351196e-14 -7.522114 9.351196e-14
## GO:0007059 chromosome segregation        2.074373e-11 -6.899759 2.074373e-11
##                                                 q.val set.size         exp1
## GO:0000279 M phase                       4.039243e-12      471 1.070282e-15
## GO:0048285 organelle fission             2.688185e-11      362 1.486831e-14
## GO:0000280 nuclear division              2.688185e-11      339 2.849163e-14
## GO:0007067 mitosis                       2.688185e-11      339 2.849163e-14
## GO:0000087 M phase of mitotic cell cycle 7.058283e-11      349 9.351196e-14
## GO:0007059 chromosome segregation        1.304781e-08      136 2.074373e-11
##
## $stats
##                                                          stat.mean     exp1
## GO:0007156 homophilic cell adhesion                       4.065745 4.065745
## GO:0016339 calcium-dependent cell-cell adhesion           3.414326 3.414326
## GO:0048729 tissue morphogenesis                           3.113452 3.113452
## GO:0002009 morphogenesis of an epithelium                 3.093930 3.093930
## GO:1901617 organic hydroxy compound biosynthetic process  2.937016 2.937016
## GO:0035295 tube development                               2.867380 2.867380
```

Reactome Analysis

```r
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]
print(paste("Total number of significant genes:", length(sig_genes)))
```

```
## [1] "Total number of significant genes: 8228"
```

```r
write.table(sig_genes, file="significant_genes.txt", row.names=FALSE, col.names=FALSE, quote=FALSE)
```

Q: What pathway has the most significant "Entities p-value"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods?

The endosomal/vacuolar pathway has the most significant p-value, almost 0! Its p-value = 8.56E-4

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] gageData_2.32.0             gage_2.44.0
##  [3] pathview_1.34.0             org.Hs.eg.db_3.14.0
##  [5] AnnotationDbi_1.56.2        DESeq2_1.34.0
##  [7] SummarizedExperiment_1.24.0 Biobase_2.54.0
##  [9] MatrixGenerics_1.6.0        matrixStats_0.61.0
## [11] GenomicRanges_1.46.0        GenomeInfoDb_1.30.0
## [13] IRanges_2.28.0              S4Vectors_0.32.2
## [15] BiocGenerics_0.40.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2             bit64_4.0.5            splines_4.1.1
##  [4] highr_0.9              blob_1.2.2             GenomeInfoDbData_1.2.7
##  [7] yaml_2.2.1             pillar_1.6.3           RSQLite_2.2.8
## [10] lattice_0.20-44        glue_1.4.2             digest_0.6.28
## [13] RColorBrewer_1.1-2     XVector_0.34.0         colorspace_2.0-2
## [16] htmltools_0.5.2        Matrix_1.3-4           XML_3.99-0.8
## [19] pkgconfig_2.0.3        genefilter_1.76.0      zlibbioc_1.40.0
## [22] GO.db_3.14.0           purrr_0.3.4            xtable_1.8-4
## [25] scales_1.1.1           BiocParallel_1.28.0    tibble_3.1.5
## [28] annotate_1.72.0        KEGGREST_1.34.0        generics_0.1.0
## [31] ggplot2_3.3.5          ellipsis_0.3.2         cachem_1.0.6
## [34] survival_3.2-11        magrittr_2.0.1         crayon_1.4.1
## [37] KEGGgraph_1.54.0       memoise_2.0.0          evaluate_0.14
## [40] fansi_0.5.0            graph_1.72.0           tools_4.1.1
## [43] lifecycle_1.0.1        stringr_1.4.0          munsell_0.5.0
## [46] locfit_1.5-9.4         DelayedArray_0.20.0    Biostrings_2.62.0
## [49] compiler_4.1.1         rlang_0.4.11           grid_4.1.1
## [52] RCurl_1.98-1.5         bitops_1.0-7           rmarkdown_2.11
## [55] gtable_0.3.0           DBI_1.1.1              R6_2.5.1
## [58] knitr_1.36             dplyr_1.0.7            fastmap_1.1.0
## [61] bit_4.0.4              utf8_1.2.2             Rgraphviz_2.38.0
## [64] stringi_1.7.5         parallel_4.1.1        Rcpp_1.0.7
## [67] vctrs_0.3.8           geneplotter_1.72.0    png_0.1-7
## [70] tidyselect_1.1.1      xfun_0.26
```